NORGES TEKNISK-NATURVITENSKAPELIGE UNIVERSITET

Sampling algorithm for estimating Bayes factor

by

Hilde G. Borgos and Henning Omre

PREPRINT
STATISTICS NO. 4/2000

NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY
TRONDHEIM, NORWAY

This report has URL http://www.math.ntnu.no/preprint/statistics/2000/S4-2000.ps
Hilde G. Borgos has homepage: http://www.math.ntnu.no/~hgb
E-mail: hgb@stat.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.

Sampling algorithm for estimating Bayes factor

Hilde G. Borgos and Henning Omre

Department of Mathematical Sciences, NTNU, Trondheim, Norway.

Abstract

Bayes factor can be used to choose between two specified models, based on available observations, without any requirement of nested models. In the case of equal prior probabilities of the two models, Bayes factor is given as the posterior odds ratio of the models. Evaluation of Bayes factor includes solving of integrals, which in most cases can only be solved numerically. This work presents an McMC algorithm constructed to estimate Bayes factor when the available observations contain sampling errors. The McMC algorithm allows simultaneous sampling of the model parameters. The posterior distribution of the unknown, high-dimensional, underlying variables is assumed to be of a form where the bulk of the variables are analytically tractable, while only a low-dimensional subset needs to be sampled from by the use of McMC. The performance of the algorithm is studied by considering a special case of exact observations where Bayes factor is analytically tractable, thus the estimates can be compared to the corresponding analytical values.

1 Introduction

In the work presented by Borgos et al. (2000), two possible distributions of the size of geological faults are studied, where the size is represented as either maximum displacement or horizontal extent of the fault. It is of interest to decide which distribution is most likely to be the correct one, based on available observations. Observations of faults are often non-exact, and knowledge of the sampling procedure should be incorporated when defining the model choice problem. In addition to the model choice, it is of interest to make inference about the underlying fault population. This inference is based on the posterior distribution of the model parameters and the fault sizes.

Bayes factor can be used as a criterion for model choice, see Kass and Raftery (1995) for an overview. In many problems encountered, Bayes factor can not be found analytically, but must be approximated using numerical integration techniques, see Evans and Swartz

(1995). Among the techniques available for numerical approximations, Markov chain Monte Carlo (McMC) sampling is well suited for dealing with complex, high-dimensional distributions, see for instance Besag et al. (1995), Han and Carlin (2000). McMC techniques can be used to estimate Bayes factor, for example by including a model indicator in the sampling, as described by Carlin and Chib (1995).

The current paper presents an McMC algorithm constructed with the aim of solving the model choice problem encountered by Borgos et al. (2000). The unknown quantities involved in this work consist of a stochastic variable, and model parameters describing the distribution of the variable under each model. The dimension of the unknown stochastic variable is high, and possibly unknown. The dimensions of the model parameters are low under both competing models, and in general need not coincide. The posterior distribution of the high-dimensional variable is assumed to be analytically tractable, and under both competing models inference about the model parameters can be based on posterior distributions obtained by an analytical integration over the high-dimensional variable. Thus the high-dimensional variable need not be included in the sample space of the McMC algorithm. General model choice problems of this kind are considered.

A Metropolis-Hastings algorithm is presented, where the target distribution is defined inspired by the target distribution of the Gibbs algorithm presented by Carlin and Chib (1995), including a model indicator as a parameter. The target distribution is constructed with the aim of estimating Bayes factor and the model parameters of the preferred model. The target distribution is not uniquely defined however, but leaves some freedom of choice that will affect neither the value of Bayes factor nor the posterior distribution of the model parameter under the preferred model. The freedom of choice in the definition of the target distribution is exploited to optimize the performance of the McMC algorithm. The estimator of Bayes factor is expressed through a statistic of the realizations from the McMC algorithm. Minimization of uncertainty of the estimator of Bayes factor is complicated, but instead minimization of the uncertainty of the statistic involved in the estimator is used as a criterion for optimality.

The general model choice problem is presented in Section 2, and Section 3 gives a discussion on estimation of Bayes factor based on McMC sampling. The model choice problem encountered by Borgos et al. (2000) is summarized in Section 4, and in Section 5 the general behavior of Bayes factor and the performance of the McMC algorithm on this particular problem is evaluated.

2 Model Choice Problem

A high-dimensional stochastic variable $\boldsymbol{Y}$ is believed to originate from one of two possible models, where the stochastic variable $M \in \Omega_m = \{m_1, m_2\} = \{1, 2\}$ indicates which model is the true one, model 1 or 2 respectively. Under model m, the probability density function (pdf) of $\boldsymbol{Y}$ is defined as $f(\boldsymbol{y}|m, \psi_m)$, where $\Psi_m \in \Omega_{\psi_m}$ is a low-dimensional stochastic vector of model parameters under model m. The dimensions of the parameters Ψ_1 and Ψ_2 under the two models need not coincide. The model parameter of model m is assumed to have a proper prior $f(\psi_m|m)$, thus the joint distribution of $\boldsymbol{Y}$ and Ψ_m is

$$f(\boldsymbol{y}, \psi_m|m) = f(\boldsymbol{y}|m, \psi_m) f(\psi_m|m). \tag{1}$$

The aim of this work is to decide which model is most likely, and to make inference about the model parameter Ψ_m corresponding to the favored model m and the variable $\boldsymbol{Y}$.

The parameter Ψ_m is unknown and the variables $\boldsymbol{Y}$ are not observed exactly, but inference can be made based on observations $\boldsymbol{X}$. The observations can include different types of errors, like precision errors, bias and censoring. After a model choice has been made, the posterior pdfs of interest are the conditional pdfs $f(\boldsymbol{y}, \psi_m|\boldsymbol{x}, m) = f(\boldsymbol{y}|\boldsymbol{x}, m, \psi_m) f(\psi_m|\boldsymbol{x}, m)$, where m is the preferred model. The observations $\boldsymbol{X}$ are described by a likelihood function $h(\boldsymbol{x}|\boldsymbol{y}, m, \psi_m)$ under model m, depending on both the high-dimensional variable $\boldsymbol{Y}$ and the low-dimensional model parameter Ψ_m. However, the likelihood function and prior distributions are assumed to be of a form where

$$h(\boldsymbol{x}|m, \psi_m) = \int h(\boldsymbol{x}|\boldsymbol{y}, m, \psi_m) f(\boldsymbol{y}|m, \psi_m)\, d\boldsymbol{y} \tag{2}$$

is analytically tractable. Thus the likelihood function can be reduced to only depending on a small set of variables. The posterior distribution of Ψ_m under model m can now be expressed as

$$f(\psi_m|\boldsymbol{x}, m) = \text{const} \times h(\boldsymbol{x}|m, \psi_m) f(\psi_m|m), \tag{3}$$

and inference about Ψ_m can be made based on an analytical randomization over $\boldsymbol{Y}$. The posterior distribution of $\boldsymbol{Y}$ under model m, conditioned on Ψ_m, is given by

$$f(\boldsymbol{y}|\boldsymbol{x}, m, \psi_m) = \frac{h(\boldsymbol{x}|\boldsymbol{y}, m, \psi_m) f(\boldsymbol{y}|m, \psi_m)}{h(\boldsymbol{x}|m, \psi_m)}. \tag{4}$$

Bayes factor can be used as a criterion for model choice when two or more alternative models are suggested, without any requirement of nested models. See Kass and Raftery (1995) for an overview of Bayes factor. In the case of two competing models, the two outcomes $m_1 = 1$ and $m_2 = 2$ of the indicator variable M are assigned prior probabilities $f(m_1)$ and $f(m_2)$, which are assumed to be known. Bayes factor is defined as the ratio between the posterior odds ratio and the prior odds ratio:

$$B_{12}(\boldsymbol{x}) = \frac{f(m_1|\boldsymbol{x})}{f(m_2|\boldsymbol{x})} \cdot \frac{f(m_2)}{f(m_1)} = \frac{f(\boldsymbol{x}|m_1)}{f(\boldsymbol{x}|m_2)}. \tag{5}$$

If there is no prior reason to prefer one model to the other, the prior distributions are chosen as $f(m_1) = f(m_2)$ and Bayes factor equals the posterior odds ratio. If $B_{12}(\boldsymbol{x}) > 1$

the observations favor model 1, while if $B_{12}(\boldsymbol{x}) < 1$ the observations support model 2. The marginal density function $f(\boldsymbol{x}|m)$ under model m is defined as

$$\begin{aligned} f(\boldsymbol{x}|m) &= \iint h(\boldsymbol{x}|\boldsymbol{y}, m, \psi_m) f(\boldsymbol{y}|m, \psi_m) f(\psi_m|m)\, d\boldsymbol{y}\, d\psi_m \\ &= \int h(\boldsymbol{x}|m, \psi_m) f(\psi_m|m)\, d\psi_m. \end{aligned} \tag{6}$$

From Expressions (5) and (6) it is observed that the value of Bayes factor does not depend on the choice of prior probabilities $f(m)$. Bayes factor can be interpreted as the ratio between two normalizing constants, since $f(\boldsymbol{x}|m)$ is given as the normalizing constant in the relationship in Expression (3):

$$f(\psi_m|\boldsymbol{x}, m) = \frac{h(\boldsymbol{x}|m, \psi_m) f(\psi_m|m)}{f(\boldsymbol{x}|m)}. \tag{7}$$

The marginal distribution in Expression (6) is analytically tractable only in specific cases. In general, Bayes factor can not be analytically obtained, but must instead be estimated. Estimation of Bayes factor is the main topic of this report.

The value of Bayes factor may depend strongly on the values of the hyper-parameters in the prior distributions of Ψ_1 and Ψ_2. As discussed by Kass and Raftery (1995), the choice of hyper-parameters can be critical to the result of the model choice. Thus strong prior information about the parameters are desirable to give a reliable conclusion of the model choice. Furthermore, only proper priors should be used. If improper priors are specified for the model parameters, Bayes factor is only found up to an unknown constant. Alternative analytical definitions of Bayes factors are suggested, giving factors less sensitive to the choice of prior distribution. Some examples are the posterior Bayes factor (Aitkin, 1991), partial Bayes factor and fractional Bayes factor (O'Hagan, 1991, 1995) and the intrinsic Bayes factor (Berger and Pericchi, 1996). A common feature of these alternative Bayes factors is that they combine an initial prior, which may be improper, with some or all of the observations to obtain a proper prior for the parameter Ψ_m. In this work the priors $f(\psi_m|m)$ are assumed to be well specified proper priors, and the original definition of Bayes factor in Expression (5) is used.

3 Estimating Bayes Factor

The integrals in Expression (6), involved in the expression of Bayes factor, are in general not analytically tractable. If no analytical solution is found, Bayes factor can be calculated by numerical integration, using for example asymptotic approximations or sampling approaches like simple Monte Carlo, importance sampling or McMC sampling. Evans and Swartz (1995) and Kass and Raftery (1995) discuss different estimation techniques. In this work the focus is on estimation techniques based on McMC sampling. See, for example, Besag et al. (1995) and references therein for an introduction to McMC theory. Han and Carlin (2000) gives an overview of McMC algorithms constructed to estimate Bayes factor.

The aim of the current work is to calculate Bayes factor in order to choose between two proposed models, and to make inference about the model parameter Ψ_m and the unknown variable $\boldsymbol{Y}$ under the favored model m, see Expressions (3) and (4). McMC techniques enable simultaneous sampling of both the model indicator M and the model parameters Ψ_1 and Ψ_2. The variable $\boldsymbol{Y}$ is not included in the McMC algorithm, but samples of the variable $\boldsymbol{Y}$ can be generated from the posterior distribution in Expression (4) after the model choice is performed and the corresponding model parameter is estimated.

Realizations from the posterior distribution of one specific model can be used to estimate Bayes factor, see Kass and Raftery (1995) for an overview. Chib (1995) and DiCiccio et al. (1997) use the fact that Bayes factor is given as a ratio between normalizing constants in posterior distributions, see Expressions (5) and (7). The normalizing constant under the specified model can be estimated based on the realizations, and by doing this for both models under consideration an estimate of Bayes factor can be obtained. Other algorithms constructed to estimate normalizing constants or ratios between normalizing constants can also be used to estimate Bayes factor, see Gelman and Meng (1998) for an overview.

Instead of considering one model at a time, the model indicator M can be included as a variable in the McMC algorithm. The Reversible Jump McMC algorithm suggested by Green (1995) generate realizations of (M, Ψ_M) from $f(m, \psi_m|\boldsymbol{x})$. The sample space of the algorithm is defined as $\Omega_m \times (\Omega_{\psi_1} \cup \Omega_{\psi_2})$. Carlin and Chib (1995) described a Gibbs sampler on the sample space $\Omega_m \times \Omega_{\psi_1} \times \Omega_{\psi_2}$, giving realizations of (M, Ψ_1, Ψ_2) from the posterior pdf $f(m, \psi_1, \psi_2|\boldsymbol{x})$. In order for this pdf to be completely defined, pdfs $\pi(\psi_{m_*}|m)$, $m_* \neq m$, must be specified. Carlin and Chib (1995) denote these pdfs pseudopriors or linking densities, which are defined under the assumption that Ψ_{m_*} is independent of Ψ_m and $\boldsymbol{X}$, given $M = m$. Dellaportas et al. (1998) suggested combining the two above mentioned algorithms. In the Gibbs algorithm of Carlin and Chib (1995), update steps of the model indicator are replaced by an acceptance/rejection step based on the Reversible Jump algorithm of Green (1995). Other sampling algorithms based on the idea of including M as a variable are presented by Carlin and Polson (1991) and George and McCulloch (1993).

3.1 McMC algorithm

In this work Bayes factor is estimated by the use of an McMC algorithm where the model indicator M is included in the sampling scheme. A target distribution $\pi(m, \psi_m, \psi_{m_*}|\boldsymbol{x})$, $m_* \neq m$, is defined, adopting the concept of pseudopriors. This pdf is defined with the aim of estimating Bayes factor and the model parameter of the favored model, and it is shown below that all pseudopriors can be chosen arbitrary without altering the quantities of interest. The target distribution is defined as

$$\begin{aligned}\pi(m, \psi_m, \psi_{m_*}|\boldsymbol{x}) &= \pi(\boldsymbol{x})^{-1} h(\boldsymbol{x}|m, \psi_m, \psi_{m_*}) f(\psi_m|m)\pi(\psi_{m_*}|m)\pi(m) \\ &= \pi(\boldsymbol{x})^{-1} h(\boldsymbol{x}|m, \psi_m) f(\psi_m|m)\pi(\psi_{m_*}|m)\pi(m). \qquad (8)\end{aligned}$$

Throughout this work the notation $\pi(\cdot)$ is used do denote pseudopriors or to stress that a pdf is defined with respect to the pseudopriors. In the likelihood function $h(\boldsymbol{x}|m, \psi_m, \psi_{m_*})$, M is an indicator of which Ψ_m applies to the observations, and the likelihood function is assumed to be independent of Ψ_{m_*}. Furthermore, Ψ_m and Ψ_{m_*} are assumed to be independent given M. Since Bayes factor is independent of the actual value of the prior pdf $f(m)$, see Expression (5), any proper discrete pdf with $\pi(m) > 0$ for all $m \in \Omega_m$ can be used as a pseudoprior for M in the target distribution (8). As pointed out by Carlin and Chib (1995), the pseudoprior $\pi(\psi_{m_*}|m)$, $m_* \neq m$, can be chosen as any proper pdf with $\pi(\psi_{m_*}|m) > 0$ for all $\psi_{m_*} \in \Omega_{\psi_{m_*}}$. This is illustrated through the expression

$$\begin{aligned}\pi(\boldsymbol{x}|m) &= \iint \pi(m)^{-1}\pi(m, \psi_m, \psi_{m_*}|\boldsymbol{x})\pi(\boldsymbol{x})\, d\psi_m\, d\psi_{m_*} \\ &= \iint h(\boldsymbol{x}|m, \psi_m) f(\psi_m|m)\pi(\psi_{m_*}|m)\, d\psi_m\, d\psi_{m_*} \\ &= \int h(\boldsymbol{x}|m, \psi_m) f(\psi_m|m)\, d\psi_m, \\ &= f(\boldsymbol{x}|m) \qquad (9)\end{aligned}$$

showing that $\pi(\boldsymbol{x}|m_1)/\pi(\boldsymbol{x}|m_2)$ gives the correct value of Bayes factor regardless of the form of the pseudopriors, see Expression (5). Furthermore, the posterior pdf

$$\begin{aligned}\pi(\psi_m|\boldsymbol{x}, m) &= \int \pi(m|\boldsymbol{x})^{-1}\pi(m, \psi_m, \psi_{m_*}|\boldsymbol{x})\, d\psi_{m_*} \\ &= \int (\pi(m|\boldsymbol{x})\pi(\boldsymbol{x}))^{-1} h(\boldsymbol{x}|m, \psi_m) f(\psi_m|m)\pi(\psi_{m_*}|m)\pi(m)\, d\psi_{m_*} \\ &= \int \pi(\boldsymbol{x}|m)^{-1} h(\boldsymbol{x}|m, \psi_m) f(\psi_m|m)\pi(\psi_{m_*}|m)\, d\psi_{m_*} \\ &= f(\boldsymbol{x}|m)^{-1}\, h(\boldsymbol{x}|m, \psi_m) f(\psi_m|m) \\ &= f(\psi_m|\boldsymbol{x}, m) \qquad (10)\end{aligned}$$

is also correct, using Expression (9), and inference about Ψ_m is not affected by the pseudopriors. Finally, the posterior distribution of $\boldsymbol{Y}$ under model m, Expression (4), is also independent of the choice of pseudopriors.

Although all marginal distributions involved in the expression of Bayes factor and the posterior distributions of interest are independent of the pseudopriors, some of the other marginal pdfs may depend on them. Some examples are:

$$\begin{aligned}\pi(m|\boldsymbol{x}) &= \pi(\boldsymbol{x})^{-1}\pi(\boldsymbol{x}|m)\pi(m) \;=\; \pi(\boldsymbol{x})^{-1} f(\boldsymbol{x}|m)\pi(m) \\ \pi(\psi_m|\boldsymbol{x}) &= f(\psi_m|\boldsymbol{x}, m)\pi(m|\boldsymbol{x}) + \pi(\psi_m|m_*)\pi(m_*|\boldsymbol{x}).\end{aligned}$$

The introduction of pseudopriors $\pi(\psi_m|m_*)$ and $\pi(\psi_{m_*}|m)$ on the model parameters is required for the posterior distribution of (M, Ψ_m, Ψ_{m_*}) to be completely specified. Furthermore, Expression (5) implies the possibility of introducing a pseudoprior $\pi(m)$ also for the model indicator, without altering the results. These pseudopriors constitute a freedom of choice in the specification of the target distribution (8), which will be exploited to minimize the uncertainty in the estimate of Bayes factor. The uncertainty of the estimate depends on the behavior of the McMC algorithm. Before optimal choices of pseudopriors can be discussed, the McMC algorithm is presented and an estimator of Bayes factor is suggested.

A Metropolis-Hastings algorithm is constructed to sample from the target distribution in Expression (8), see Hastings (1970). The algorithm involves sampling of the model indicator M and the model parameters Ψ_1 and Ψ_2, conditioned on observations $\boldsymbol{X}$. Three transition steps are defined within one iteration of the algorithm, one for each variable. The McMC algorithm for estimating Bayes factor based on observations $\boldsymbol{X}$ can be expressed as follows:

Algorithm 1

Let $\Omega = \Omega_m \times \Omega_{\psi_m} \times \Omega_{\psi_{m_*}}$ be the sample space of (M, Ψ_m, Ψ_{m_*}) and assume the target distribution in Expression (8) satisfies $\pi(m, \psi_m, \psi_{m_*}|\boldsymbol{x}) > 0; \forall (m, \psi_m, \psi_{m_*}) \in \Omega$.

- *Initiate*
  - *Arbitrary $(m(0), \psi_m(0), \psi_{m_*}(0)) \in \Omega$.*
- *Iterate, $t = 1, 2, \ldots$*
  - Step 1*:*
    - *Denote $m(t-1) = m$, $\psi_m(t-1) = \psi_m$ and $\psi_{m_*}(t-1) = \psi_{m_*}$.*
    - *Propose m' from $q(m'|m)$.*
    - *Calculate*
      for $m' = m_* \neq m$:

$$\begin{aligned}\alpha(m_*|m) \quad = \quad \min\Bigg\{1, \frac{\pi(m_*)q(m|m_*)}{\pi(m)q(m_*|m)} \\ \times \frac{h(\boldsymbol{x}|m_*, \psi_{m_*})f(\psi_{m_*}|m_*)}{\pi(\psi_{m_*}|m)} \times \frac{\pi(\psi_m|m_*)}{h(\boldsymbol{x}|m, \psi_m)f(\psi_m|m)}\Bigg\},\end{aligned} \tag{11}$$

      for $m' = m$: $\alpha(m|m) = 1$
    - *Let: $m(t) = m'$ with probability $\alpha(m'|m)$,*
      $m(t) = m$ else.

– Step 2*:*

  * *Denote* $m(t) = m$ *and* $\psi_m(t-1) = \psi_m$.
  * *Propose* ψ'_m *from* $q(\psi'_m|\psi_m, m)$.
  * *Calculate*

$$\alpha(\psi'_m|\psi_m, m) \;=\; \min\left\{1, \frac{h(\boldsymbol{x}|m, \psi'_m)f(\psi'_m|m)q(\psi_m|\psi'_m, m)}{h(\boldsymbol{x}|m, \psi_m)f(\psi_m|m)q(\psi'_m|\psi_m, m)}\right\}. \qquad (12)$$

  * *Let:* $\psi_m(t) = \psi'_m$ *with probability* $\alpha(\psi'_m|\psi_m, m)$,
    $\psi_m(t) = \psi_m$ *else.*

– Step 3*:*

  * *Denote* $m(t) = m$ *and* $\psi_{m_*}(t-1) = \psi_{m_*}$ *where* $m_* \neq m$.
  * *Propose* ψ'_{m_*} *from* $q(\psi'_{m_*}|\psi_{m_*}, m)$.
  * *Calculate*

$$\alpha(\psi'_{m_*}|\psi_{m_*}, m) \;=\; \min\left\{1, \frac{\pi(\psi'_{m_*}|m)q(\psi_{m_*}|\psi'_{m_*}, m)}{\pi(\psi_{m_*}|m)q(\psi'_{m_*}|\psi_{m_*}, m)}\right\}. \qquad (13)$$

  * *Let:* $\psi_{m_*}(t) = \psi'_{m_*}$ *with probability* $\alpha(\psi'_{m_*}|\psi_{m_*}, m)$,
    $\psi_{m_*}(t) = \psi_{m_*}$ *else.*

The variables $(M(t), \Psi_m(t), \Psi_{m_*}(t))$ at iteration t of the algorithm have marginal pdf $\pi^t(m, \psi_m, \psi_{m_*}|\boldsymbol{x})$ where $\lim_{t\to\infty} \pi^t(m, \psi_m, \psi_{m_*}|\boldsymbol{x}) = \pi(m, \psi_m, \psi_{m_*}|\boldsymbol{x})$. The systematic scan of the three steps of the algorithm does not imply a time-reversible Markov chain, but the chain is both aperiodic and irreducible. Lack of time-reversibility has no effect on the estimates presented below. Time-reversibility can be obtained by drawing at random the order of the steps at each iteration. Note that if the pseudoprior is used as transition density for Ψ_{m_*}, $q(\psi'_{m_*}|\psi_{m_*}, m) = \pi(\psi'_{m_*}|m)$, the acceptance probability (13) becomes $\alpha(\psi'_{m_*}|\psi_{m_*}, m) = 1$ and the change is always accepted.

Both Bayes factor $B_{12}(\boldsymbol{x})$ and the model parameters Ψ_1 and Ψ_2 can be estimated based on realizations obtained from McMC simulations. Suppose that n subsequent samples $\{m^j, \psi_1^j, \psi_2^j; j = 1, \ldots, n\}$ are gathered after convergence of the algorithm is reached. Estimation of $B_{12}(\boldsymbol{x})$ is essentially the same as estimating the posterior pdf $\pi(m_1|\boldsymbol{x})$, and a natural estimator of the latter is given by the relative number of outcomes of model 1:

$$\hat{\pi}(m_1|\boldsymbol{x}) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}[M^j = m_1]. \qquad (14)$$

The marginal distribution of M^j is $\pi(m|\boldsymbol{x})$, thus the mean and variance of the estimator are given as

$$
\begin{aligned}
\mathrm{E}\left[\hat{\pi}(m_1|\boldsymbol{x})\right] &= \frac{1}{n}\sum_{j=1}^{n}\mathrm{E}\left[\mathbf{1}[M^j=m_1]\right] = \pi(m_1|\boldsymbol{x}) \\
\mathrm{Var}\left[\hat{\pi}(m_1|\boldsymbol{x})\right] &= \frac{1}{n^2}\sum_{j=1}^{n}\mathrm{Var}\left[\mathbf{1}[M^j=m_1]\right] + \frac{2}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n-j}\mathrm{Cov}[\mathbf{1}[M^j=m_1],\mathbf{1}[M^{j+k}=m_1]] \\
&= \frac{1}{n}\pi(m_1|\boldsymbol{x})(1-\pi(m_1|\boldsymbol{x})) + \frac{2}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n-j}\gamma(k), \qquad (15)
\end{aligned}
$$

where $\gamma(k) = \mathrm{Cov}[\mathbf{1}[M^j = m_1], \mathbf{1}[M^{j+k} = m_1]]$ is the autocovariance at lag k. A natural estimator of $B_{12}(\boldsymbol{x})$, see Expression (5), is given as

$$
\hat{B}_{12}(\boldsymbol{x}) = \frac{\hat{\pi}(m_1|\boldsymbol{x})}{1-\hat{\pi}(m_1|\boldsymbol{x})}\cdot\frac{\pi(m_2)}{\pi(m_1)}. \qquad (16)
$$

Recall that from the definition of $B_{12}(\boldsymbol{x})$, its true value is independent of the pseudoprior $\pi(m)$. However, the distribution of the estimator $\hat{B}_{12}(\boldsymbol{x})$ depends on the choice of $\pi(m)$. Note that the mean and variance of $\hat{B}_{12}(\boldsymbol{x})$

$$
\begin{aligned}
\mathrm{E}\left[\hat{B}_{12}(\boldsymbol{x})\right] &= \frac{\pi(m_2)}{\pi(m_1)}\mathrm{E}\left[\frac{\hat{\pi}(m_1|\boldsymbol{x})}{1-\hat{\pi}(m_1|\boldsymbol{x})}\right] \\
\mathrm{Var}\left[\hat{B}_{12}(\boldsymbol{x})\right] &= \left(\frac{\pi(m_2)}{\pi(m_1)}\right)^2\mathrm{Var}\left[\frac{\hat{\pi}(m_1|\boldsymbol{x})}{1-\hat{\pi}(m_1|\boldsymbol{x})}\right] \qquad (17)
\end{aligned}
$$

need not be finite, since $\hat{\pi}(m_1|\boldsymbol{x})$ may equal 1.

An estimator of the parameter value in model m is given by the empirical posterior mean

$$
\hat{\Psi}_m = \hat{\mathrm{E}}\left[\Psi_m|\boldsymbol{x},m\right] = \frac{\sum_{j=1}^{n}\Psi_m^j\mathbf{1}[M^j=m]}{\sum_{j=1}^{n}\mathbf{1}[M^j=m]}. \qquad (18)
$$

The subset $\{\psi_m^j|m^j = m\}$ of realizations from the McMC algorithm can also be used to describe the complete posterior pdf $f(\psi_m|\boldsymbol{x}, m)$. Finally, realizations of $\boldsymbol{Y}$ under the favored model m can be obtained by sampling from the conditional posterior distribution $f(\boldsymbol{y}|\boldsymbol{x}, m, \psi_m)$ given in Expression (4). An example is given in Section 4.2, where it is explained how $f(\boldsymbol{y}|\boldsymbol{x}, m, \psi_m)$ can be sampled from by the use of rejection sampling.

3.2 Pseudopriors

The pseudopriors $\pi(\psi_1|m_2)$, $\pi(\psi_2|m_1)$ and $\pi(m)$ can be chosen as any proper pdfs satisfying $\pi(\psi_m|m_*) > 0; \forall\psi_m \in \Omega_{\psi_m}$, $m = m_1, m_2$, and $\pi(m) > 0; \forall m \in \Omega_m$. In search for good pseudopriors, the major goal is to reduce uncertainty in the estimator of Bayes factor. The pseudopriors are chosen focusing on the behavior of M in the McMC algorithm. In addition, the transition densities $q(\psi'_m|\psi_m, m)$ and $q(\psi'_{m_*}|\psi_{m_*}, m)$ in Algorithm 1 should be chosen to span the sample space of the model parameters Ψ_1 and Ψ_2 adequately.

The estimator (16) of Bayes factor need not have a finite mean and variance, see Expression (17), and this complicates evaluation of the properties of the estimator. The precision of $\hat{B}_{12}(\boldsymbol{x})$ is influenced by the properties of the estimator $\hat{\pi}(m_1|\boldsymbol{x})$, see Expression (14), and minimization of the uncertainty of the latter estimator is considered. If the variables M^j are independent, $\gamma(k) = 0$ for $k \neq 0$ in Expression (15), and $\mathrm{Var}\,[\hat{\pi}(m_1|\boldsymbol{x})] = n^{-1}\pi(m_1|\boldsymbol{x})(1 - \pi(m_1|\boldsymbol{x}))$. However, the variance of of $\hat{\pi}(m_1|\boldsymbol{x})$ is further reduced if the autocovariance at lag 1 is negative, $\gamma(1) < 0$, giving a negative correlation between succeeding realizations of M. The extreme case is $\gamma(1) = -\mathrm{Var}\,[\mathbf{1}[M = m_1]] = -\pi(m_1|\boldsymbol{x})(1 - \pi(m_1|\boldsymbol{x}))$, where M alternates periodically between m_1 and m_2, a situation that arises if the probability of a move from m to m' in one step of Algorithm 1 satisfies:

$$\mathrm{P}\{M(t) = m'|M(t-1) = m\} = q(m'|m)\alpha(m'|m) = \begin{cases} 0, & m' = m \\ 1, & m' = m_* \neq m. \end{cases} \tag{19}$$

If this extreme case should actually be obtained, the periodicity of the Markov chain can be removed through minor adjustments of $q(m'|m)$. However, in practice the requirement will never be exactly reached, and periodicity is no problem. Although probably never obtained, this extreme case still serves as a target for optimal achievement. The requirement (19) can only be fulfilled if

$$\alpha(m_*|m) = 1, \quad q(m'|m) = \mathbf{1}[m' \neq m] = \begin{cases} 0, & m' = m \\ 1, & m' = m_* \neq m. \end{cases}$$

Thus the acceptance probability in Expression (11) should satisfy

$$\alpha(m_*|m) = \min\left\{1, \frac{\pi(m_*)}{\pi(m)} \cdot \frac{h(\boldsymbol{x}|m_*, \psi_{m_*})f(\psi_{m_*}|m_*)}{\pi(\psi_{m_*}|m)} \cdot \frac{\pi(\psi_m|m_*)}{h(\boldsymbol{x}|m, \psi_m)f(\psi_m|m)}\right\} = 1, \tag{20}$$

while $\alpha(m|m) = 1$ always holds in Algorithm 1. The two last ratios in $\alpha(m_*|m)$ are independent of the model parameters if their pseudopriors are chosen equal to the posterior pdfs,

$$\pi(\psi_1|m_2) = f(\psi_1|\boldsymbol{x}, m_1), \qquad \pi(\psi_2|m_1) = f(\psi_2|\boldsymbol{x}, m_2). \tag{21}$$

The two ratios are then given as

$$\frac{h(\boldsymbol{x}|m_*, \psi_{m_*})f(\psi_{m_*}|m_*)}{f(\psi_{m_*}|\boldsymbol{x}, m_*)} \cdot \frac{f(\psi_m|\boldsymbol{x}, m)}{h(\boldsymbol{x}|m, \psi_m)f(\psi_m|m)} = \frac{f(\boldsymbol{x}|m_*)}{f(\boldsymbol{x}|m)} = B_{m_*m}(\boldsymbol{x}), \tag{22}$$

resulting in the requirement

$$\alpha(m_*|m) = \min\left\{1, \frac{\pi(m_*)}{\pi(m)} B_{m_*m}(\boldsymbol{x})\right\} = 1, \tag{23}$$

which is fulfilled if the pseudoprior of M is chosen as

$$\pi(m_1) = (1 + B_{12}(\boldsymbol{x}))^{-1}. \tag{24}$$

Thus the requirement in Expression (19) is obtained if the transition density is chosen as $q(m'|m) = \mathbf{1}[m' \neq m]$ and the pseudopriors are chosen as given in Expressions (21) and (24). In this case the variance of the estimator $\hat{\pi}(m_1|\boldsymbol{x})$ is minimized. The fulfillment of the requirement in Expression (23) implies $\pi(m|\boldsymbol{x}) = \pi(m_*|\boldsymbol{x}) = 0.5$, which coincides with the estimate $\hat{\pi}(m|\boldsymbol{x}) = 0.5$ obtained when M alternates periodically between m and m_*.

The optimal choice of pseudopriors is expressed through Bayes factor and the posterior pdfs of the model parameters, which are the unknown quantities the algorithm is constructed to estimate. Thus an optimal algorithm can only be constructed if the answers are known beforehand, in which case there would be no need for the McMC sampling. However, the knowledge of the optimal choice can serve as a guideline for how to choose the pseudopriors, by choosing pdfs sufficiently close to the optimal solution. One way of doing this is to run an McMC algorithm sampling independently from $f(\psi_1|\boldsymbol{x}, m_1)$ and $f(\psi_2|\boldsymbol{x}, m_2)$, and use the realizations of Ψ_1 and Ψ_2 to estimate normalized posterior pdfs and to give a rough estimate of Bayes factor. This sampling can be performed by Algorithm 1, using only the transition kernel $q(\psi'_m|\psi_m, m)$. If r realizations Ψ^j_m, $j = 1, \ldots, r$ are generated, the pseudoprior of Ψ_m can be chosen as a kernel density estimator

$$\pi(\psi_m|m_*) = \hat{f}(\psi_m|\boldsymbol{x}, m) = \frac{1}{r}\sum_{j=1}^{r} g(\psi_m|\psi^j_m, h_m), \tag{25}$$

where the kernel $g(\psi_m|\psi^j_m, h_m)$ is a proper pdf parameterized by ψ^j_m and h_m, see Silverman (1986). By including the observations ψ^j_m in the parameterization, the kernel densities can be scaled to give normalized densities on Ω_{ψ_m}, and the pseudoprior is defined as a proper pdf conditioned on $\Psi_m \in \Omega_{\psi_m}$. The pseudoprior of M in Expression (24) is based on a preliminary estimate of Bayes factor, given by

$$\tilde{B}_{12}(\boldsymbol{x}) = \frac{1}{L}\sum_{j=1}^{L} \frac{h(\boldsymbol{x}|m_1, \tilde{\psi}^j_1) f(\tilde{\psi}^j_1|m_1)}{\pi(\tilde{\psi}^j_1|m_2)} \cdot \frac{\pi(\tilde{\psi}^j_2|m_1)}{h(\boldsymbol{x}|m_2, \tilde{\psi}^j_2) f(\tilde{\psi}^j_2|m_2)}, \tag{26}$$

where $\tilde{\psi}^j_m$ are arbitrary values of Ψ_m, picked from the set of realizations. This is a similar approach as used by Chib (1995), who emphasize that normalizing constants can be estimated by evaluating the likelihood function, prior pdf and estimated posterior pdf only for one parameter value. From Expression (22) it is observed that if the optimal choices of pseudopriors $\pi(\psi_m|m_*)$ are used, Expression (21), the estimate $\tilde{B}_{12}(\boldsymbol{x})$ will give an exact solution regardless of the values of $\tilde{\psi}^j_m$. If the pseudopriors $\pi(\psi_m|m_*)$ are sufficiently close to the posterior pdfs, the rough estimate (26) can be based on a single value $\tilde{\psi}_m$, i.e., $L = 1$. In this case, when the pseudopriors from Expression (25) are used the estimate of the posterior pdf is assumed to be best near the mode of $\pi(\psi_m|m_*)$, and $\tilde{\psi}_m$ should be chosen from this area.

The choice of pseudopriors aims at obtaining negatively correlated realizations of M, to minimize the uncertainty in the estimator $\hat{\pi}(m_1|\boldsymbol{x})$. Since the optimal choice of pseudopriors can not be reached, the extreme case with $q(m'|m) = \mathbf{1}[m' \neq m]$ and $\alpha(m'|m) = 1$ is not fulfilled. The transition density $q(m'|m)$ is left unchanged, but the acceptance probability for a change of model will in practice depend on both the model M and the parameter values ψ_m and ψ_{m_*}. Note that the negative correlation obtained when using optimal pseudopriors only applies to M, and not the model parameters Ψ_m and Ψ_{m_*}. The correlation structure for the model parameters depend on the transition kernels $q(\psi'_m|\psi_m, m)$ and $q(\psi'_{m_*}|\psi_{m_*}, m)$ in Algorithm 1, which are left unspecified in the discussion above.

The McMC algorithm is run twice, first to adjust pseudopriors, next to estimate Bayes factor and the model parameter of the favored model. Note, however, that although both runs of the algorithm use the same observations, the limiting distributions of interest in the model choice problem are not influenced by this repeated use of the data. The first McMC algorithm sampling from the posterior distributions is only used to improve the performance of the algorithm by adjusting the pseudopriors and exploiting the freedom in choosing these pdfs. In the second and final run of the McMC algorithm, the limiting distributions involved in the expression of Bayes factor and the posterior pdfs of the model parameters are unaffected by the pseudopriors, see Expressions (9) and (10).

3.3 Uncertainty in estimated Bayes factor

In general the acceptance probability $\alpha(m'|m)$ is influenced by the parameter values ψ_m and ψ_{m_*}, thus it is not straightforward to quantify the uncertainty in the estimated Bayes factor. In the case of $\{M^j; j = 1, \ldots, n\}$ being independent realizations from the posterior pdf $\pi(m|\boldsymbol{x})$, an approximate confidence interval for $B_{12}(\boldsymbol{x})$ can be obtained. Then $\mathrm{P}\{M^j = m_1\} = \pi(m_1|\boldsymbol{x}) = p$ and $n\hat{\pi}(m_1|\boldsymbol{x})$ follows a binomial distribution $n\hat{\pi}(m_1|\boldsymbol{x}) \sim \mathrm{bin}(n, p)$, see Expression (14), which for large n can be approximated by a normal distribution $n\hat{\pi}(m_1|\boldsymbol{x}) \sim N(np, np(1-p))$. From Expression (5) p can be expressed as a function of $B_{12}(\boldsymbol{x})$, and combined with the normal approximation this is used to find an approximate $(1-\alpha)100\%$ confidence interval for $B_{12}(\boldsymbol{x})$:

$$\hat{B}_{12}(\boldsymbol{x}) + \frac{\pi(m_2)}{\pi(m_1)} \cdot \frac{u_{\alpha/2}^2}{2n(1-\hat{\pi}(m_1|\boldsymbol{x}))^2} \left(1 \pm \left(1 + \frac{4n\hat{\pi}(m_1|\boldsymbol{x})(1-\hat{\pi}(m_1|\boldsymbol{x}))}{u_{\alpha/2}^2} \right)^{1/2} \right), \tag{27}$$

where $u_{\alpha/2}$ is a critical value of the standard normal distribution. The interval is not centered at $\hat{B}_{12}(\boldsymbol{x})$, but it is easily shown that the estimator is covered by the interval.

If the pseudopriors are sufficiently well adapted so that negatively correlated samples of M are generated, the approximate confidence interval will be a conservative interval with too wide confidence limits. If on the other hand samples of M are positively correlated, the estimated confidence interval obtained under the assumption of independent samples is likely to give too narrow confidence limits. In order to reduce the correlation, realizations used in the interval estimation should be separated by a number of iterations, on the expense of having fewer samples. The separation length depends on the mixing of the algorithm. Although not all samples are used to find a confidence interval for Bayes factor, the best point estimate $\hat{B}_{12}(\boldsymbol{x})$ is still obtained by using all samples.

4 Example: Geological Faults

In the work by Borgos et al. (2000) two models are suggested for the distribution of the size of geological faults. The fault sizes of k faults are represented by a vector $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_k)$ of iid stochastic variables, where size is measured as maximum displacement, and under model m the pdf of $\boldsymbol{Y}$ is given by

$$f(\boldsymbol{y}|m, k, \theta_m) = \prod_{j=1}^{k} f(y_j|m, \theta_m). \tag{28}$$

The number of faults, K, is treated as a stochastic variable in cases where also the dimension of $\boldsymbol{Y}$ is unknown. A power law, or fractal, distribution is frequently used to model fault displacements, see for example JSG (1996, Vol. 18). This fractal model corresponds to the Pareto distribution,

$$f(y|m_1, \theta_1) = \frac{\theta_1 y_0^{\theta_1}}{y^{\theta_1+1}}, \qquad y \geq y_0,\ \theta_1 > 0, \tag{29}$$

which is suggested as the distribution of Y under model 1. The pareto distribution is described by two parameters, Θ_1 and y_0, where the lower limit $y_0 > 0$ must be defined in order for the pdf $f(y|m_1, \theta_1) \propto y^{-(\theta_1+1)}$ of the power law distribution to be a proper pdf. The sample space of Θ_1 is $\Omega_{\theta_1} = (0, \infty)$, while the lower limit y_0 is assumed to be known. As a competing model, a truncated exponential distribution is suggested:

$$f(y|m_2, \theta_2) = \theta_2 \exp(-\theta_2(y - y_0)), \qquad y \geq y_0,\ \theta_2 > 0. \tag{30}$$

The exponential distribution is described by the parameter Θ_2, with sample space $\Omega_{\theta_2} = (0, \infty)$. Exponential distributions of fault displacements have been observed, for example, on mid-ocean ridges, see Cowie et al. (1994), and in clay models, see Spyropoulos et al. (1999) and Ackermann et al. (1999). A common lower bound $y_0 > 0$ is used for the two competing distributions. This enables calculation of Bayes factor based on observations where the limit of resolution has imposed a lower bound on observable values.

It is of interest to estimate true fault sizes $\boldsymbol{Y}$ based on observations $\boldsymbol{X}$, including estimation of K if the dimension of $\boldsymbol{Y}$ is unknown. Two alternative sampling regimes are considered. In Section 4.1 the case of exact observations is discussed, while the more general case of non-exact observations is discussed in Section 4.2. In Section 4.3, Algorithm 1 is adapted to the model choice problem for fault size distribution.

4.1 Exact observations

The simplest sampling situation is the case of exact observations $\boldsymbol{X} \equiv \boldsymbol{Y}$, where all k variables are observed without any measurement errors. The likelihood function is defined through a Dirac delta function $\delta(\cdot)$:

$$h(\boldsymbol{x}|\boldsymbol{y}, m, \theta_m) = \prod_{j=1}^{k} \delta(y_j - x_j), \tag{31}$$

thus with $\Psi_m = \Theta_m$ the likelihood function in Expression (2) becomes

$$h(\boldsymbol{x}|m, \theta_m) = \prod_{j=1}^{k} f(x_j|m, \theta_m). \tag{32}$$

Under the two suggested models in Expressions (29) and (30), this likelihood function can be expressed as follows:

$$\begin{aligned} h(\boldsymbol{x}|m_1, \theta_1) &= (\prod_{j=1}^{k} x_j)^{-1} \cdot \theta_1^k \exp\{-\theta_1 \sum_{j=1}^{k} (\ln x_j - \ln x_0)\}, \\ h(\boldsymbol{x}|m_2, \theta_2) &= \theta_2^k \exp\{-\theta_2 \sum_{j=1}^{k} (x_j - x_0)\}, \end{aligned}$$

where $x_0 = y_0$. Both densities belong to an exponential family, and are given on the form

$$h(\boldsymbol{x}|m, \theta_m) = c_m(\boldsymbol{x})\, \theta_m^k \exp\{-\theta_m t_m(\boldsymbol{x})\}, \tag{33}$$

where the functions $c_m(\boldsymbol{x})$ and $t_m(\boldsymbol{x})$ are given as $c_1(\boldsymbol{x}) = (\prod_{j=1}^{k} x_j)^{-1}$, $c_2(\boldsymbol{x}) = 1$, $t_1(\boldsymbol{x}) = \sum_{j=1}^{k} (\ln x_j - \ln x_0)$ and $t_2(\boldsymbol{x}) = \sum_{j=1}^{k} (x_j - x_0)$. From Expression (6) the marginal pdf of $\boldsymbol{X}$ under model m is given as

$$f(\boldsymbol{x}|m) = c_m(\boldsymbol{x}) \int \theta_m^k \exp\{-\theta_m t_m(\boldsymbol{x})\} f(\theta_m|m)\, d\theta_m. \tag{34}$$

This pdf can be found analytically if conjugate priors of the exponential family are chosen, for example gamma distributions with hyper-parameters α_m and β_m, with pdf

$$f(\theta_m|m) = \Gamma(\alpha_m)^{-1} \beta_m^{-\alpha_m} \theta_m^{\alpha_m - 1} \exp(-\theta_m/\beta_m), \qquad \theta_m \geq 0,\ \alpha_m, \beta_m > 0 \tag{35}$$

and mean and variance $\mathrm{E}\left[\Theta_m|m\right] = \alpha_m \beta_m$, $\mathrm{Var}\left[\Theta_m|m\right] = \alpha_m \beta_m^2$. The marginal pdf in (34) is then found to be

$$\begin{aligned} f(\boldsymbol{x}|m) &= c_m(\boldsymbol{x}) \Gamma(\alpha_m)^{-1} \beta_m^{-\alpha_m} \int_0^{\infty} \theta_m^{\alpha_m + k - 1} \exp\{-\theta_m(\beta_m^{-1} + t_m(\boldsymbol{x}))\}\, d\theta_m \\ &= c_m(\boldsymbol{x}) \Gamma(\alpha_m)^{-1} \Gamma(\alpha_m + k) \beta_m^{-\alpha_m} (\beta_m^{-1} + t_m(\boldsymbol{x}))^{-(\alpha_m + k)}, \end{aligned} \tag{36}$$

and Bayes factor is given as

$$B_{12}(\boldsymbol{x}) = (\prod_{j=1}^{k} x_j)^{-1} \frac{\beta_2^{\alpha_2}(\beta_2^{-1} + \sum_{j=1}^{k}(x_j - x_0))^{\alpha_2+k}}{\beta_1^{\alpha_1}(\beta_1^{-1} + \sum_{j=1}^{k}(\ln x_j - \ln x_0))^{\alpha_1+k}} \cdot \frac{\Gamma(\alpha_2)\Gamma(\alpha_1+k)}{\Gamma(\alpha_1)\Gamma(\alpha_2+k)}. \tag{37}$$

The posterior distribution of Θ_m under model m is a gamma distribution with parameters $(\alpha_m + k)$ and $(\beta_m^{-1} + t_m(\boldsymbol{x}))^{-1}$, and mean and variance

$$\mathrm{E}\left[\Theta_m|\boldsymbol{x}, m\right] = \frac{\beta_m(\alpha_m + k)}{1 + \beta_m t_m(\boldsymbol{x})}, \qquad \mathrm{Var}\left[\Theta_m|\boldsymbol{x}, m\right] = \frac{\beta_m^2(\alpha_m + k)}{(1 + \beta_m t_m(\boldsymbol{x}))^2}. \tag{38}$$

Similar calculations can be executed using an ordinary exponential distribution with no lower bound, replacing the function $t_2(\boldsymbol{x})$ with $t_2(\boldsymbol{x}) = \sum_{j=1}^{k} x_j$.

4.2 Non-exact observations

Observations of fault sizes are in general often contaminated with measurement errors. Furthermore, only a limited number $N \leq K$ of the faults are observed. For each variable Y_j there is a probability $p(Y_j)$ of observing the fault. The vector $\boldsymbol{S} = (S_1, S_2, \ldots, S_K)$ of indicator variables describes which variables that are observed, where $S_j = 1$ if variable Y_j is observed and $S_j = 0$ otherwise. The number of observed variables is thus $N = \sum_{j=1}^{K} S_j$. If a variable Y_j is observed, the likelihood of the measured value, denoted Z_j, is described by the function $h(z_j|y_j)$. The complete set of observations is denoted $\boldsymbol{X} = (\boldsymbol{S}, \boldsymbol{Z})$.

The unknown variable K has the same interpretation under both models, being the total number of faults. However, the posterior pdfs of K can differ significantly between the two models, and instead of handling a problem of bimodality of K in the posterior, auxiliary variables K_1 and K_2 are introduced to represent the number of faults under the two models. The unknown dimension K_m of $\boldsymbol{Y}$ under model m is thus included in the model parameter $\Psi_m = (K_m, \Theta_m)$. Furthermore, depending on how prior knowledge about K_m has been obtained, the prior pdfs of K_m may be chosen differently under the two models. This is discussed by Borgos et al. (2000). The sample spaces of K_1 and K_2 are given as $\Omega_{k_1} = \Omega_{k_2} = \{0, 1, 2, \ldots\}$. However, since both K_1 and K_2 must satisfy $K_m \geq N$, given the observations, the likelihood function $h(\boldsymbol{x}|\boldsymbol{y}, m, \psi_m)$ has value zero for $K_m < N$. For $K_m \geq N$, the likelihood function of the observations $\boldsymbol{X}$ is given as

$$h(\boldsymbol{x}|\boldsymbol{y}, m, \psi_m) = \prod_{j=1}^{k_m} p(y_j)^{s_j}(1 - p(y_j))^{1-s_j} \prod_{j;s_j=1} h(z_j|y_j). \tag{39}$$

The reduced likelihood function in Expression (2) is found to be

$$\begin{aligned} h(\boldsymbol{x}|m,\psi_m) &= \prod_{j;s_j=1} \int h(z_j|y_j)p(y_j)f(y_j|m,\theta_m)\,dy_j \prod_{j;s_j=0} \int (1-p(y_j))f(y_j|m,\theta_m)\,dy_j \\ &= (1-p(m,\theta_m))^{k_m-n} \prod_{i=1}^{n} h(z_i|m,\theta_m), \end{aligned} \tag{40}$$

where

$$p(m,\theta_m) = \int p(y)f(y|m,\theta_m)\,dy, \qquad h(z|m,\theta_m) = \int h(z|y)p(y)f(y|m,\theta_m)\,dy. \tag{41}$$

For simplicity, the observed values $\boldsymbol{z}$ are renumbered from $\{j; s_j = 1\}$ to $i = 1, \ldots, n$, where $n = \sum_{j=1}^{k_m} s_j$. In the work by Borgos et al. (2000) the probability $p(y)$ is an monotonely increasing, piecewise linear function and $h(z|y)$ is a uniform pdf on a closed interval determined by y, thus the integrals in Expression (41) are analytically tractable.

The model choice problem is solved using the McMC algorithm described below. When a model m has been chosen and the corresponding model parameters $\Psi_m = (K_m, \Theta_m)$ estimated, realizations of $\boldsymbol{Y}$ from $f(\boldsymbol{y}|\boldsymbol{x}, m, \psi_m)$, Expression (4), can be obtained by rejection sampling. The posterior pdf of $\boldsymbol{Y}$ can be expressed as

$$f(\boldsymbol{y}|\boldsymbol{x},m,\psi_m) = \prod_{j;s_j=1} \frac{h(z_j|y_j)p(y_j)f(y_j|m,\theta_m)}{h(z_j|m,\theta_m)} \prod_{j;s_j=0} \frac{(1-p(y_j))f(y_j|m,\theta_m)}{1-p(m,\theta_m)},$$

see Expression (40). The pdf $f(y_j|m,\theta_m)$ can be used as a proposal distribution for all j, and the sample y_j is accepted with probability $h(z_j|y_j)p(y_j)/\max_y\{h(z_j|y)p(y)\}$ if $s_j = 1$ and with probability $(1-p(y_j))/\max_y\{1-p(y)\}$ if $s_j = 0$. For the piecewise linear function $p(y)$ and the uniform pdf $h(z|y)$ used by Borgos et al. (2000), the maximum values are easily obtained.

4.3 McMC algorithm

Algorithm 1 is used to solve the model choice problem for fault size distributions, and to estimate the parameters of the preferred model. The algorithm is constructed for the general case of non-exact observations. The target distribution in Expression (8) is now given as

$$\begin{aligned} &\pi(m,\psi_m,\psi_{m_*}|\boldsymbol{x}) = \\ &\qquad \text{const} \times (1-p(m,\theta_m))^{k_m-n} \prod_{i=1}^{n} h(z_i|m,\theta_m) f(k_m,\theta_m|m)\pi(k_{m_*},\theta_{m_*}|m)\pi(m), \end{aligned} \tag{42}$$

where $\boldsymbol{X} = (\boldsymbol{S}, \boldsymbol{Z})$ and $\Psi_m = (K_m, \Theta_m)$. The pseudoprior of Ψ_m, based on realizations $\{\psi_m^j\}$, $j = 1, \ldots, r$ from the posterior pdf $f(\psi_m|\boldsymbol{x}, m)$, is given by the kernel density estimate

$$\pi(\psi_m|m_*) = \frac{1}{r}\sum_{j=1}^{r} g_k(k_m|k_m^j, \lambda_m)g_\theta(\theta_m|\theta_m^j, \sigma_m). \tag{43}$$

A truncated Gaussian kernel with variance σ_m^2 is used for Θ_m, conditioned on $\Theta_m \in \Omega_{\theta_m}$. A Poisson distribution with parameter λ_m is used for the absolute value of $K_m - k_m^j$, with equal probability of positive and negative difference. If $k_m < k_m^j$, conditioning on $K_m \in \Omega_{k_m}$ is included in the kernel density. The parameters σ_m and λ_m are chosen as

$$\sigma_m = (4/3)^{1/5}\,\widehat{\mathrm{std}}(\Theta_m)\,r^{-1/5}, \qquad \lambda_m = (4/3)^{1/5}\,\widehat{\mathrm{std}}(K_m)\,r^{-1/5},$$

see Silverman (1986), where the standard deviations of Θ_m and K_m are estimated based on the realizations $\{\psi_m^j\}$, $j = 1, \ldots, r$.

There are three update steps within one iteration of Algorithm 1, updating M, Ψ_m and Ψ_{m_*} separately. When M is updated, the transition density is chosen as $q(m'|m) = \mathbf{1}[m' \neq m]$, thus $m' = m_*$ is always suggested. The acceptance probability is given in Expression (11), where $q(m_*|m) = q(m|m_*) = 1$ cancels. New values of Θ_m and K_m are suggested independently, using a transition kernel

$$q(\psi_m'|\psi_m, m) = q(\theta_m'|\theta_m, m)q(k_m'|k_m, m). \tag{44}$$

A new value θ_m' is suggested from a Gaussian distribution with mean θ_m and variance $\tau_m = 0.05\,\mathrm{E}\,[\Theta_m|m]$, truncated at $\theta_m \pm 5\tau_m$. Thus $q(\theta_m'|\theta_m, m) = q(\theta_m|\theta_m', m)$. If $\theta_m' \leq 0$, i.e., $\theta_m' \notin \Omega_{\theta_m}$, the acceptance probability in Expression (12) is replaced by $\alpha(\psi_{m_*}'|\psi_{m_*}, m) = 0$ and the suggested values $\psi_m' = (\theta_m', k_m')$ are not accepted, regardless of k_m and k_m'.

The current number k_m is suggested changed to either $k_m + 1$ or $k_m - 1$, choosing the first alternative with probability $\beta(k_m)$, the second with probability $1 - \beta(k_m)$. Since $K_m \geq n$, the function value $\beta(n)$ is chosen as $\beta(n) = 1$. An increase from k_m to $k_m + 1$ corresponds to adding a new element to the vector $\boldsymbol{Y}$ of faults, where the new element has $k_m + 1$ possible index numbers. A decrease from k_m to $k_m - 1$ corresponds to removing one of the unobserved faults, that is, removing a Y_j where $S_j = 0$. There are then $k_m - n$ possible indexes j to choose from. The two alternative transition kernels of K_m are thus given as

$$q(k_m + 1|k_m, m) = \frac{\beta(k_m)}{k_m + 1}, \qquad q(k_m - 1|k_m, m) = \frac{1 - \beta(k_m)}{k_m - n}.$$

When $\theta_m' > 0$, the acceptance probability for the suggested change in Ψ_m is given in Expression (12). The pseudoprior (43) can easily be sampled from, and is used as transition kernel for Ψ_{m_*}. Thus $q(\psi_{m_*}'|\psi_{m_*}, m) = \pi(\psi_{m_*}'|m)$ and the acceptance probability in Expression (13) is $\alpha(\psi_{m_*}'|\psi_{m_*}, m) = 1$.

The McMC algorithm can easily be simplified to deal with the case of exact observations. The likelihood function from Expression (40) is replaced by the function in Expression (32), and the number of faults, K, is no longer treated as a stochastic variable. This has implications on the pseudoprior and the transition kernels of Ψ_m and Ψ_{m_*}, where all terms involving K are removed.

5 Evaluation of the McMC Algorithm

The behavior of Bayes factor and the performance of Algorithm 1 are examined by considering the special case of a Pareto distribution and an exponential distribution as competing models, see Section 4. Bayes factor can be found analytically in the case of exact observations, which permits a comparison of the estimated Bayes factor with the analytical value in Expression (37). The posterior means of the model parameters can be estimated based on samples from the McMC algorithm, see Expression (18), and can be compared to the exact posterior means in Expression (38).

Consider the case of exact observations. In some of the simulation studies presented below a true model m is chosen, with a corresponding model parameter θ_m, and a set of observations $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ is generated from $f(\boldsymbol{x}|m, \theta_m)$ using the distributions presented in Section 4. Bayes factor is then calculated for a selection of values of the hyper-parameters in the prior pdf of Θ_{m_*}, $m_* \neq m$. Gamma priors are used, Expression (35), with a common constant shape parameter $\alpha_1 = \alpha_2 = 4$ for both parameters Θ_1 and Θ_2. For the true model the scale parameter is chosen as $\beta_m = \theta_m/\alpha_m$, thus the chosen parameter value coincides with the prior mean $\mathrm{E}\left[\Theta_m|m\right] = \alpha_m\beta_m$. For the other model the scale parameter is given as $\beta_{m_*} = \mathrm{E}\left[\Theta_{m_*}|m_*\right]/\alpha_{m_*}$, where a number of values of the prior mean $\mathrm{E}\left[\Theta_{m_*}|m_*\right]$ is used to explore Bayes factor. Before the simulation study is carried out, the range of values of $\mathrm{E}\left[\Theta_{m_*}|m_*\right]$ to be used was determined by a preliminary simulation study.

To study the convergence of Algorithm 1, both the convergence of Ψ_m and and the mixing of M should be explored. The convergence of the algorithm can be examined by plotting different functions of the variables for realizations from the simulations. Several runs of the algorithm should be compared, using different initial states, to reveal any dependence on the starting point. The burn-in period of the algorithm is used to secure convergence of Ψ_m to the posterior distribution under model m, and M is of no interest. During these iterations independent simulations of Ψ_m and Ψ_{m_*} are performed with their respective posterior distributions as target distributions, using transition kernels $q(\psi'_m|\psi_m, m)$ and $q(\psi'_{m_*}|\psi_{m_*}, m_*)$. After the burn-in a continued sampling of Ψ_m and Ψ_{m_*} from their respective posterior distributions is performed, giving realizations from which the pseudopriors can be defined. Finally, the McMC algorithm is run with initial states $\psi_m(0)$ chosen as the final realizations of the preliminary runs and $m(0)$ uniformly drawn from Ω_m. The mixing properties of the Markov chain is then examined based on the realizations of the model indicator M.

In Section 5.1 the behavior of Bayes factor is studied through randomization of $\boldsymbol{X}$, while in Section 5.2 the performance of Algorithm 1 is explored in the case of exact observations, using a single dataset $\boldsymbol{X}$ generated from $f(\boldsymbol{x}|m, \theta_m)$. In Section 5.3 an example with non-exact observations is presented.

5.1 Behavior of analytical Bayes factor for exact observations

The behavior of the analytical Bayes factor in the case of exact observations, see Expression (37), is studied by generating 25 000 realizations of the vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)$ from the Pareto distribution with parameter $\Theta_1 = 1.0$ and lower limit $x_0 = 2$. For each set of observations, Bayes factor is calculated for a selection of prior mean values $\mathrm{E}[\Theta_2|m_2]$. Model 1 is known to be the correct model, thus Bayes factor should be $B_{12} > 1$. However, due to natural variations in the generated samples, the wrong model, model 2, in some cases fits the data better than the true model. Table 1 reports the results for sample sizes of $k = 10$ and $k = 100$, listing the mean values of Bayes factor based on the 25 000 samples. The geometric mean and the median are also reported. The empirical probability $P(B_{12} < 1)$ gives the proportion of samples where model 2 is favored. Within the range of prior mean values of Θ_2 reported in Table 1, Bayes factor can be $B_{12} < 1$ for as much as $25\% - 30\%$ of the realizations for sample sizes of $k = 10$. As the size of the dataset increases, the proportion of wrong conclusions decreases, and for $k = 100$ Bayes factor favors model 1 in allmost all simulations.

Table 2 lists the result of a similar simulation study, where 25 000 samples of size $k = 10$ and $k = 100$ are generated from the exponential distribution with parameter $\Theta_2 = 0.2$ and lower limit $x_0 = 2$. Bayes factor is evaluated for different values of the prior mean $\mathrm{E}[\Theta_1|m_1]$.

The results of Tables 1 and 2 show that for exact observations from either the Pareto or the exponential distribution, there is usually no doubt which model is the right one, even for fairly small datasets. In the case of non-exact observations however, there may be greater uncertainties about which model is most likely to be the correct one.

5.2 Performance of the algorithm for exact observations

The performance of Algorithm 1 in the case of exact observations is studied by generating a dataset $\boldsymbol{X}$ consisting of $k = 10$ observations from model 1, a Pareto distribution with parameter $\Theta_1 = 1.0$ and lower limit $x_0 = 2$. For this particular dataset, the posterior mean of Θ_1 is $\mathrm{E}[\Theta_1|\boldsymbol{x}, m_1] = 0.732$, see Expression (38), while the prior mean is $\mathrm{E}[\Theta_1|m_2] = 1.0$. Inference is based on realizations from the algorithm, repeating the procedure using

		$\mathrm{E}[\Theta_2 \mid m_2]$				
		0.05	0.10	0.20	0.30	0.40
	arithmetic mean	$\mathcal{O}(10^{41})$	$\mathcal{O}(10^{42})$	$\mathcal{O}(10^{43})$	$\mathcal{O}(10^{44})$	$\mathcal{O}(10^{45})$
B_{12}	geometric mean	564.36	100.52	69.80	97.61	151.76
	median	200.77	18.26	5.67	6.04	9.12
$\hat{P}(B_{12} < 1)$		0.037	0.129	0.248	0.282	0.268

		$\mathrm{E}[\Theta_2 \mid m_2]$				
		0.05	0.10	0.15	0.20	0.25
	arithmetic mean	$\mathcal{O}(10^{300})$	$\mathcal{O}(10^{301})$	$\mathcal{O}(10^{302})$	$\mathcal{O}(10^{302})$	$\mathcal{O}(10^{303})$
B_{12}	geometric mean	$\mathcal{O}(10^{37})$	$\mathcal{O}(10^{37})$	$\mathcal{O}(10^{37})$	$\mathcal{O}(10^{37})$	$\mathcal{O}(10^{37})$
	median	$\mathcal{O}(10^{29})$	$\mathcal{O}(10^{28})$	$\mathcal{O}(10^{28})$	$\mathcal{O}(10^{29})$	$\mathcal{O}(10^{29})$
$\hat{P}(B_{12} < 1)$		0	0.00012	0.00020	0.00024	0.00024

Table 1: Variation in Bayes factor for 25 000 realizations generated from the Pareto distribution with parameter $\Theta_1 = 1.0$, using different prior means of the parameter Θ_2 in the competing exponential distribution. In the upper table $k = 10$, in the lower table $k = 100$. Arithmetic mean, geometric mean and median of the exact values of Bayes factor are reported, together with the proportions of Bayes factors of values below 1. $\mathcal{O}(x)$ gives the order of magnitude.

		$\mathrm{E}[\Theta_1 \mid m_1]$				
		0.5	0.75	1.00	1.25	1.50
	arithmetic mean	0.176	0.376	0.488	0.516	0.494
B_{12}	geometric mean	0.067	0.120	0.133	0.123	0.105
	median	0.061	0.112	0.125	0.116	0.098
$\hat{P}(B_{12} < 1)$		0.978	0.939	0.917	0.915	0.920

		$\mathrm{E}[\Theta_1 \mid m_1]$				
		0.85	0.95	1.00	1.05	1.10
	arithmetic mean	0.00538	0.00622	0.00651	0.00672	0.00687
B_{12}	geometric mean	$\mathcal{O}(10^{-8})$	$\mathcal{O}(10^{-8})$	$\mathcal{O}(10^{-8})$	$\mathcal{O}(10^{-8})$	$\mathcal{O}(10^{-8})$
	median	$\mathcal{O}(10^{-8})$	$\mathcal{O}(10^{-8})$	$\mathcal{O}(10^{-8})$	$\mathcal{O}(10^{-8})$	$\mathcal{O}(10^{-8})$
$\hat{P}(B_{12} < 1)$		0.9998	0.9998	0.9998	0.9998	0.9998

Table 2: Variation in Bayes factor for 25 000 realizations generated from the exponential distribution with parameter $\Theta_2 = 0.2$, using different prior means of the parameter Θ_1 in the competing Pareto distribution. In the upper table $k = 10$, in the lower table $k = 100$. Arithmetic mean, geometric mean and median of the exact values of Bayes factor are reported, together with the proportions of Bayes factors of values below 1. $\mathcal{O}(x)$ gives the order of magnitude.

different values of $\mathrm{E}[\Theta_2|m_2]$. The algorithm is first run to generate realizations of Θ_m, $m = 1, 2$, from their respective posterior distributions $f(\theta_m|\boldsymbol{x}, m)$. These realizations are used to define pseudopriors, and each realization is separated by 10 iterations, after a burn-in of 5000 iterations. The last realizations of Θ_1 and Θ_2 are used as initial states $\theta_1(0)$ and $\theta_2(0)$ in the final run of Algorithm 1, while an initial state $m(0)$ is chosen uniformly on Ω_m. The number of iterations is 50 000. All realizations of M are used to estimate Bayes factor and to find a confidence interval, while inference about the model parameters are based on every $10th$ realization of (Θ_1, Θ_2).

The convergence of Θ_1 and Θ_2 during a burn-in period is illustrated in Figure 1, using $\mathrm{E}[\Theta_2|m_2] = 0.1$. The figure shows realizations of Θ_m and the logarithmic values of the pdfs $f(\boldsymbol{x}|m, \theta_m)$ from 5000 iterations. Four different starting points for each parameter are used, chosen as $10^{-6}\,\mathrm{E}[\Theta_m|m]$, $\mathrm{E}[\Theta_m|m]$, $10\,\mathrm{E}[\Theta_m|m]$ and $25\,\mathrm{E}[\Theta_m|m]$. The posterior mean is close to the prior mean for both model parameters in the case studied in Figure 1, thus with the prior means as initial states convergence is soon established. Starting close to zero, the McMC algorithm also converges rapidly. With a starting point an order of magnitude larger than the posterior mean, a larger number of iterations is required before the chain converges, but in both examples convergence is clearly reached within 5000 iterations. The plots of $\log f(\boldsymbol{x}|m, \theta_m)$ support the interpretation of the convergence. The speed of convergence in the case of exact observations depend on the size of the update steps, defined through the transition kernel $q(\theta'_m|\theta_m, m)$, see Section 4.3.

Figure 2 shows an output from the algorithm, based on 50 000 iterations, with $\mathrm{E}[\Theta_2|m_2] = 0.1$. 5000 burn-in iterations are used, and the pseudopriors are based on 5000 realizations of Θ_1 and Θ_2, separated by 10 iterations. The numerical values of the estimates from this run of the algorithm are reported below, in Table 3. The plots of $\pi(m, \theta_1, \theta_2|\boldsymbol{x})$, θ_1 and θ_2 all seem stable without any trends, see Figure 2a–c. Based on the plot of the convergence of the estimate of Bayes factor, Figure 2d, it seems that in this case 50 000 iterations is sufficient to give a reliable estimate. The samples of M are not plotted, but from the acceptance rate and autocorrelation given in Table 3 it is clear that the mixing of M is satisfactory. The histograms of Θ_m in Figure 2e,f show that in this example the posterior pdfs are narrower than the prior pdfs, but the posterior pdfs are not considerably skewed compared to the priors. Figures 2g,h show the function

$$S(x|m, \theta_m) = \mathrm{P}\{X \geq x|m, \theta_m\} = \int_x^\infty f(x|m, \theta_m)\,dx, \tag{45}$$

using the posterior mean estimates as model parameters, together with the empirical estimate

$$\hat{S}(x|m, \theta_m) = \frac{1}{k}\sum_{j=1}^{k} \mathbf{1}[x_j \geq x].$$

Under model 1 the transformation $(\log x, \log S(x|m, \theta_m))$ gives a linear plot, while under model 2 linearity is obtained for $(x, \log S(x|m, \theta_m))$. The comparisons of the parametric

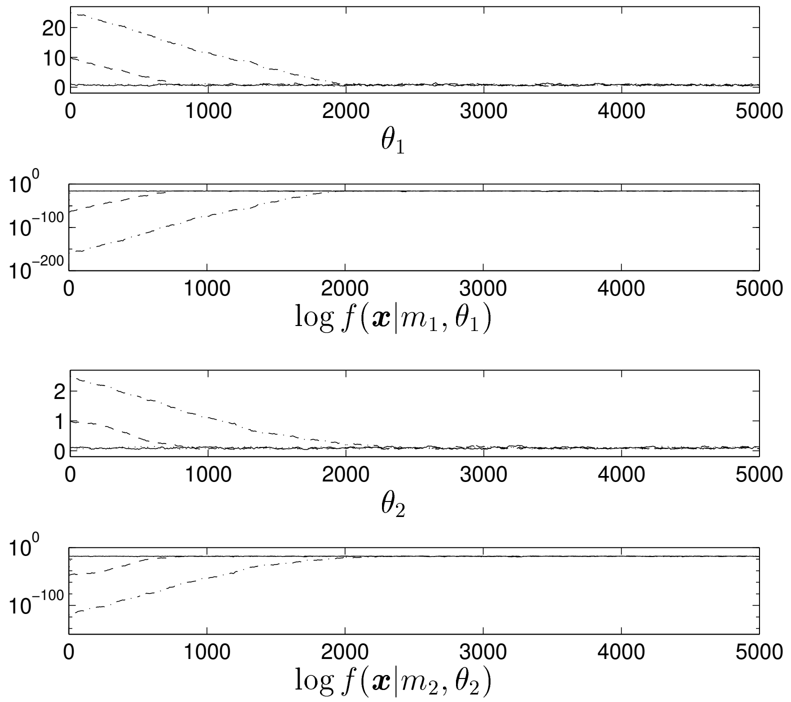


Figure 1: Realizations of Θ_1 and Θ_2 and the corresponding $\log f(\boldsymbol{x}|m,\theta_m)$ from 5000 burn-in iterations of Algorithm 1. Solid lines correspond to runs with initial values $\theta_m(0) = \mathrm{E}\left[\Theta_m|m\right]$. Dotted lines have initial values close to zero and are hardly distinguishable from the solid lines. For the dashed lines and dashed/dotted lines the initial values are $10\,\mathrm{E}\left[\Theta_m|m\right]$ and $25\,\mathrm{E}\left[\Theta_m|m\right]$ respectively.

estimates of $S(x|m,\theta_m)$ and the empirical estimates give a visual illustration of how well the observations are fitted by each of the two models.

Repeated simulations are run with a varying value of $\mathrm{E}\left[\Theta_2|m_2\right]$, and the results are listed in Table 3. The preliminary estimate $\tilde{B}_{12}(\boldsymbol{x})$ in Expression (26) based on $L=1$ parameter value, the final estimate $\hat{B}_{12}(\boldsymbol{x})$ from Expression (16), and an approximate 95% confidence interval, Expression (27), are compared with the analytical value of Bayes factor. The acceptance rate of proposed changes of model and the estimated autocorrelation $\rho(1)$ at lag 1 are also reported. Estimated posterior mean values $\hat{\mathrm{E}}\left[\Theta_m|\boldsymbol{x},m\right]$ of the model parameters, Expression (18), are compared to the true posterior means. The estimated standard deviation of the estimator $\hat{\mathrm{E}}\left[\Theta_m|\boldsymbol{x},m\right]$ is given in brackets, obtained as $(\hat{\mathrm{Var}}\left[\Theta_m|\boldsymbol{x},m\right]/n_m)^{1/2}$, where n_m is the number of realizations from model m, $\{\theta_m^j|m^j=m\}$.

For the dataset used in this simulations study, Bayes factor obtains a minimum as a function of $\beta_2 = \mathrm{E}\left[\Theta_2|m_2\right]/\alpha_2$ when $\mathrm{E}\left[\Theta_2|m_2\right] \approx 0.10$. When $\mathrm{E}\left[\Theta_2|m_2\right]$ is close to this value Bayes factor is $B_{12} < 1$, which means that model 2 actually fits the data better. From the results in Table 1, it is clear that this is not a rare event for small sample sizes. As $\mathrm{E}\left[\Theta_2|m_2\right]$

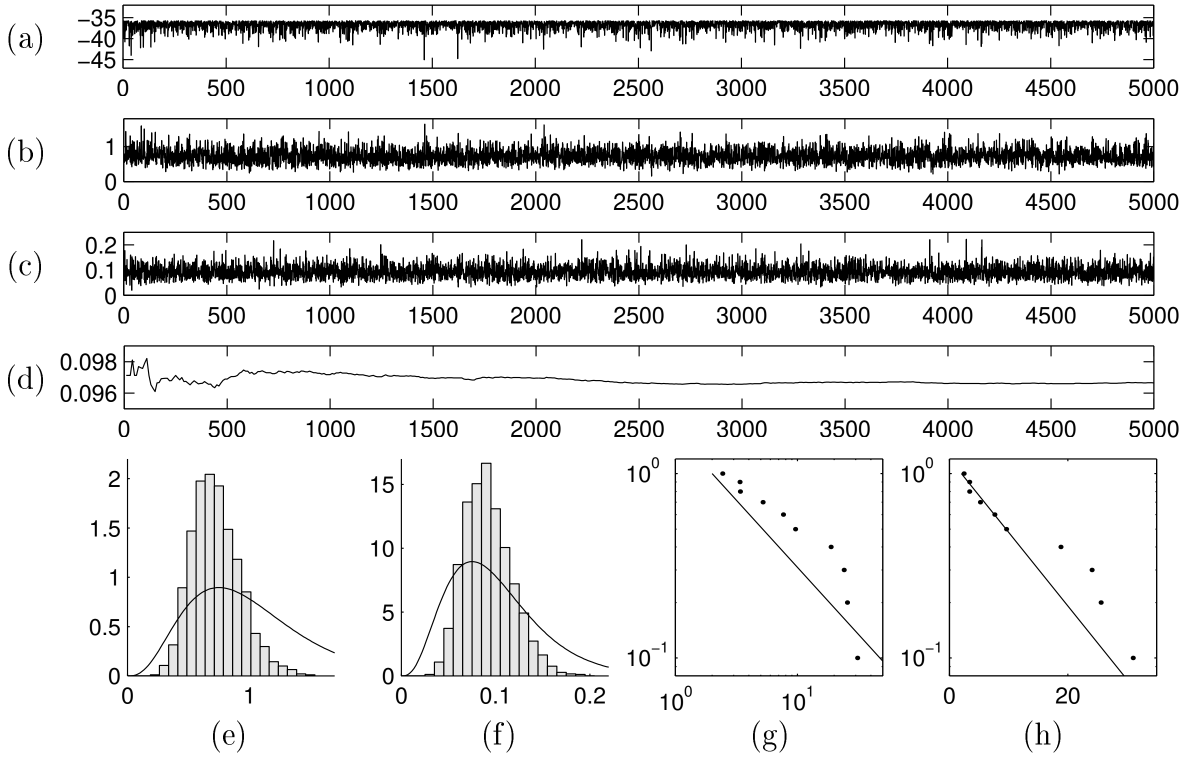
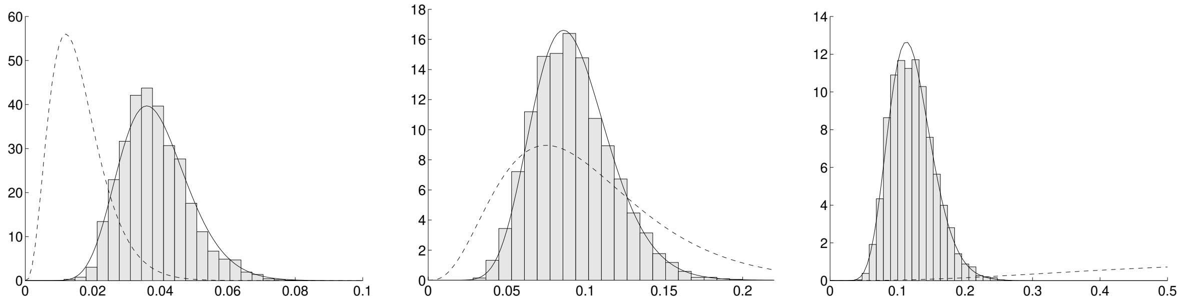


Figure 2: Output from a run of Algorithm 1 with exact observations. 5000 samples of (M, θ_1, θ_2) are generated from $\pi(m, \theta_1, \theta_2|\boldsymbol{x})$, separated by 10 iterations. The first plots are trace plots of (a) $\log \pi(m, \theta_1, \theta_2|\boldsymbol{x})$, (b) θ_1, (c) θ_2 and (d) $\hat{B}_{12}(\boldsymbol{x})$. Histograms of (e) $\{\theta_1|m = m_1\}$ and (f) $\{\theta_2|m = m_2\}$ are plotted, comparing with the corresponding prior pdfs (solid lines). The last two figures shows $S(x|m, \hat{\theta}_m)$ (solid lines), see Expression (45), together with empirical estimates. The functions are plotted on log-log scale under model 1, figure (g), and with logarithmic vertical axis under model 2, figure (h).



$\mathrm{E}[\Theta_2|m_2] = 0.016$ $\mathrm{E}[\Theta_2|m_2] = 0.10$ $\mathrm{E}[\Theta_2|m_2] = 1.0$

$\mathrm{E}[\Theta_2|m_2, \boldsymbol{x}] = 0.0388$ $\mathrm{E}[\Theta_2|m_2, \boldsymbol{x}] = 0.0927$ $\mathrm{E}[\Theta_2|m_2, \boldsymbol{x}] = 0.122$

Figure 3: Histograms of 5000 samples from the posterior distribution of $[\Theta_2|m_2]$, compared with the posterior pdf (solid line) and prior pdf (dashed lines).

deviates from 0.1, the analytical value of Bayes factor increases and eventually $B_{12} > 1$ and the correct model is favored. The estimate of Bayes factor captures this increase both for increasing and decreasing values of $\mathrm{E}[\Theta_2|m_2]$, and the analytical Bayes factor is covered by the estimated confidence interval in all simulations reported in Table 3. The estimated autocorrelation shows a clear negative correlation between succeeding realizations of M, thus the estimated confidence intervals are likely to be too conservative. The estimated posterior mean values of the parameters lie within two estimated standard deviations from the exact posterior mean in most cases, and Figure 3 shows how the histograms of realizations of $[\Theta_2|m_2]$ coincide with the posterior pdf. A comparison with the prior pdf is also included in the figure.

A similar simulation study is run by generating a sample of size $k = 10$ from model 2, an exponential distribution with parameter $\Theta_2 = 0.1$ and lower limit $x_0 = 2$. Bayes factor for this dataset is found, varying the prior mean $\mathrm{E}[\Theta_1|m_1]$, and the results are listed in Table 4. Once more the algorithm seems to give satisfactory results of both estimated Bayes factor and model parameters.

The results reported in Tables 3 and 4 illustrate the performance of the algorithm in the simplest case of exact observations and a single model parameter. Only two datasets are presented in the simulation study above, one from each distribution (29) and (30), but repeated simulations are expected to give similar results. Small datasets are chosen to illustrate the performance of the algorithm, since Bayes factor is then likely to have a value not too far from 1, see Tables 1 and 2. In this situation it is important to be able to estimate Bayes factor accurately, to give the right conclusion of the model choice problem.

$\mathrm{E}[\Theta_2 \mid m_2]$	$B_{12}(\boldsymbol{x})$	$\tilde{B}_{12}(\boldsymbol{x})$	$\hat{B}_{12}(\boldsymbol{x})$	95% CI	rate	$\hat{\rho}(1)$
0.010	249.222	290.983	248.174	(243.848,252.577)	0.895	-0.80
0.016	12.609	13.930	12.652	(12.432,12.876)	0.929	-0.86
0.025	1.358	1.311	1.356	(1.332,1.380)	0.945	-0.89
0.040	0.268	0.263	0.268	(0.263,0.272)	0.954	-0.91
0.063	0.116	0.112	0.115	(0.113,0.118)	0.971	-0.94
0.10	0.097	0.093	0.097	(0.095,0.098)	0.964	-0.93
0.16	0.147	0.151	0.148	(0.145,0.150)	0.970	-0.94
0.25	0.336	0.339	0.335	(0.329,0.341)	0.970	-0.94
0.40	1.120	1.122	1.119	(1.100,1.139)	0.976	-0.95
0.63	4.489	4.197	4.499	(4.421,4.579)	0.961	-0.92
1.0	21.475	20.396	21.465	(21.092,21.844)	0.967	-0.93
1.6	117.124	125.897	117.438	(115.396,119.517)	0.961	-0.92

$\mathrm{E}[\Theta_2 \mid m_2]$	$\mathrm{E}[\Theta_2 \mid \boldsymbol{x}, m_2]$	$\hat{\mathrm{E}}[\Theta_2 \mid \boldsymbol{x}, m_2]$	$\hat{\mathrm{E}}[\Theta_1 \mid \boldsymbol{x}, m_1]$
0.010	0.0274	0.0274 (0.0001)	0.723 (0.004)
0.016	0.0388	0.0386 (0.0002)	0.731 (0.004)
0.025	0.0516	0.0515 (0.0003)	0.744 (0.004)
0.040	0.0663	0.0666 (0.0004)	0.731 (0.004)
0.063	0.0802	0.0803 (0.0004)	0.733 (0.004)
0.10	0.0927	0.0919 (0.0005)	0.727 (0.004)
0.16	0.1029	0.1025 (0.0005)	0.739 (0.004)
0.25	0.1102	0.1102 (0.0006)	0.731 (0.004)
0.40	0.1156	0.1149 (0.0006)	0.730 (0.004)
0.63	0.1192	0.1192 (0.0007)	0.733 (0.004)
1.0	0.1216	0.1222 (0.0007)	0.728 (0.004)
1.6	0.1233	0.1236 (0.0006)	0.724 (0.004)

Table 3: McMC results for a dataset of size $k = 10$ generated from the Pareto distribution with parameter $\Theta_1 = 1.0$, varying the prior mean of Θ_2. The exact value of $B_{12}(\boldsymbol{x})$ from Expression (37) is compared to the preliminary estimate $\tilde{B}_{12}(\boldsymbol{x})$, Expression (26), the final estimate $\hat{B}_{12}(\boldsymbol{x})$, Expression (16), and the confidence interval in Expression (27). The acceptance rate of suggested changes of model and the estimated autocorrelation $\rho(1)$ at lag 1 is reported. Estimates of model parameters are given with standard deviation of the estimates in brackets. In all simulations the prior and posterior mean of Θ_1 are $\mathrm{E}[\Theta_1 \mid m_1] = 1.0$ and $\mathrm{E}[\Theta_1 \mid \boldsymbol{x}, m_1] = 0.732$.

$\mathrm{E}[\Theta_1\|m_1]$	$B_{12}(\boldsymbol{x})$	$\tilde{B}_{12}(\boldsymbol{x})$	$\hat{B}_{12}(\boldsymbol{x})$	95% CI	rate	$\hat{\rho}(1)$
0.10	0.000147	0.000139	0.000148	(0.000145,0.000151)	0.883	-0.77
0.16	0.00347	0.00361	0.00346	(0.00340,0.00352)	0.899	-0.80
0.25	0.0395	0.0430	0.0394	(0.0387,0.0401)	0.917	-0.84
0.40	0.253	0.309	0.252	(0.248,0.257)	0.888	-0.80
0.63	0.734	0.758	0.734	(0.721,0.747)	0.940	-0.88
1.00	1.075	1.032	1.074	(1.055,1.093)	0.948	-0.90
1.60	0.841	0.863	0.842	(0.827,0.857)	0.909	-0.82
2.50	0.416	0.398	0.417	(0.409,0.424)	0.952	-0.91
4.00	0.137	0.143	0.137	(0.135,0.139)	0.966	-0.93
6.30	0.0364	0.0355	0.0364	(0.0358,0.0371)	0.951	-0.90
10.00	0.00793	0.00763	0.00792	(0.00778,0.00806)	0.950	-0.90
16.00	0.00149	0.00141	0.00150	(0.00147,0.00152)	0.949	-0.90

$\mathrm{E}[\Theta_1\|m_1]$	$\mathrm{E}[\Theta_1\|\boldsymbol{x},m_1]$	$\hat{\mathrm{E}}[\Theta_1\|\boldsymbol{x},m_1]$	$\hat{\mathrm{E}}[\Theta_2\|\boldsymbol{x},m_2]$
0.10	0.282	0.280 (0.002)	0.1618 (0.0009)
0.16	0.405	0.403 (0.002)	0.1631 (0.0008)
0.25	0.547	0.550 (0.003)	0.1628 (0.0009)
0.40	0.714	0.713 (0.004)	0.1620 (0.0008)
0.63	0.877	0.874 (0.005)	0.1613 (0.0009)
1.00	1.029	1.027 (0.006)	0.1625 (0.0009)
1.60	1.156	1.151 (0.006)	0.1622 (0.0009)
2.50	1.249	1.252 (0.006)	0.1626 (0.0009)
4.00	1.319	1.321 (0.007)	0.1627 (0.0008)
6.30	1.366	1.373 (0.007)	0.1627 (0.0009)
10.00	1.398	1.398 (0.008)	0.1611 (0.0008)
16.00	1.420	1.426 (0.008)	0.1628 (0.0009)

Table 4: McMC results for a dataset of size $k = 10$ generated from the exponential distribution with parameter $\Theta_2 = 0.1$, varying the prior mean of Θ_1. The exact value of $B_{12}(\boldsymbol{x})$ from Expression (37) is compared to the preliminary estimate $\tilde{B}_{12}(\boldsymbol{x})$, Expression (26), the final estimate $\hat{B}_{12}(\boldsymbol{x})$, Expression (16), and the confidence interval in Expression (27). The acceptance rate of suggested changes of model and the estimated autocorrelation $\rho(1)$ at lag 1 is reported. Estimates of model parameters are given with standard deviation of the estimates in brackets. In all simulations the prior and posterior mean of Θ_2 are $\mathrm{E}[\Theta_2|m_2] = 0.1$ and $\mathrm{E}[\Theta_2|\boldsymbol{x}, m_2] = 0.1626$.

5.3 Performance of the algorithm for non-exact observations

When observations of fault sizes contain sampling errors, and the dimension K of the variable $\boldsymbol{Y}$ is unknown, K is included in the McMC sampling. Borgos et al. (2000) analyze a fault size dataset from the Gullfaks field, see Fossen and Rørnes (1996), and Algorithm 1 is used to estimate Bayes factor to decide if the data is best fitted by a Pareto or an exponential distribution. The observed number of faults is $n = 169$, while the total number is unknown. The convergence during burn-in iterations is first studied, and Figure 4 shows a convergence plot for the total number of faults under the two competing models, using four different initial states $k_m(0)$. Under both models, convergence is reached within 10 000 iterations. In this example the dimensions K_m under the two models are assigned equal prior distributions, Poisson distributions with mean 720. It is observed from Figure 4 that the two Markov chains stabilizes at different values, thus the posterior pdfs of K_m differ. Convergence plots for the model parameters are not included for these simulations, but also these parameters are found to reach convergence within 10 000 iterations.

Figure 5 shows an output from the McMC algorithm, based on 50 000 iterations. The numerical results of this and similar runs for the same dataset are presented by Borgos et al. (2000). A burn-in of 10 000 iterations is used, and pseudopriors are based on 5000 realizations of Ψ_1 and Ψ_2, separated by 10 iterations. In this case the preliminary and final estimate of Bayes factor, Expressions (26) and (16), are $\tilde{B}_{12}(\boldsymbol{x}) = 0.01423$ and $\hat{B}_{12}(\boldsymbol{x}) = 0.00538$, the acceptance rate for suggested changes of M is 0.0136 and the estimated autocorrelation at lag 1 is $\hat{\rho}(1) = 0.966$. The rough estimate is based on $L = 100$ values ψ_m^j.

Based on the difference between the final estimate of Bayes factor and the rough estimate, it seems that the rough estimate is not as good as the case was for exact observations with only one unknown model parameter for each model. This also means that the pseudopriors $\pi(\psi_m|m_*)$ deviates to some extent from the corresponding posterior pdfs. This

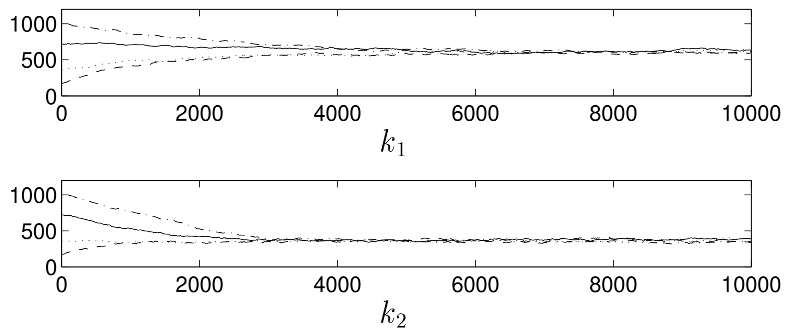


Figure 4: Realizations of K_1 and K_2 from 10 000 burn-in iterations of Algorithm 1, starting with different initial states $k_1(0)$ and $k_2(0)$.

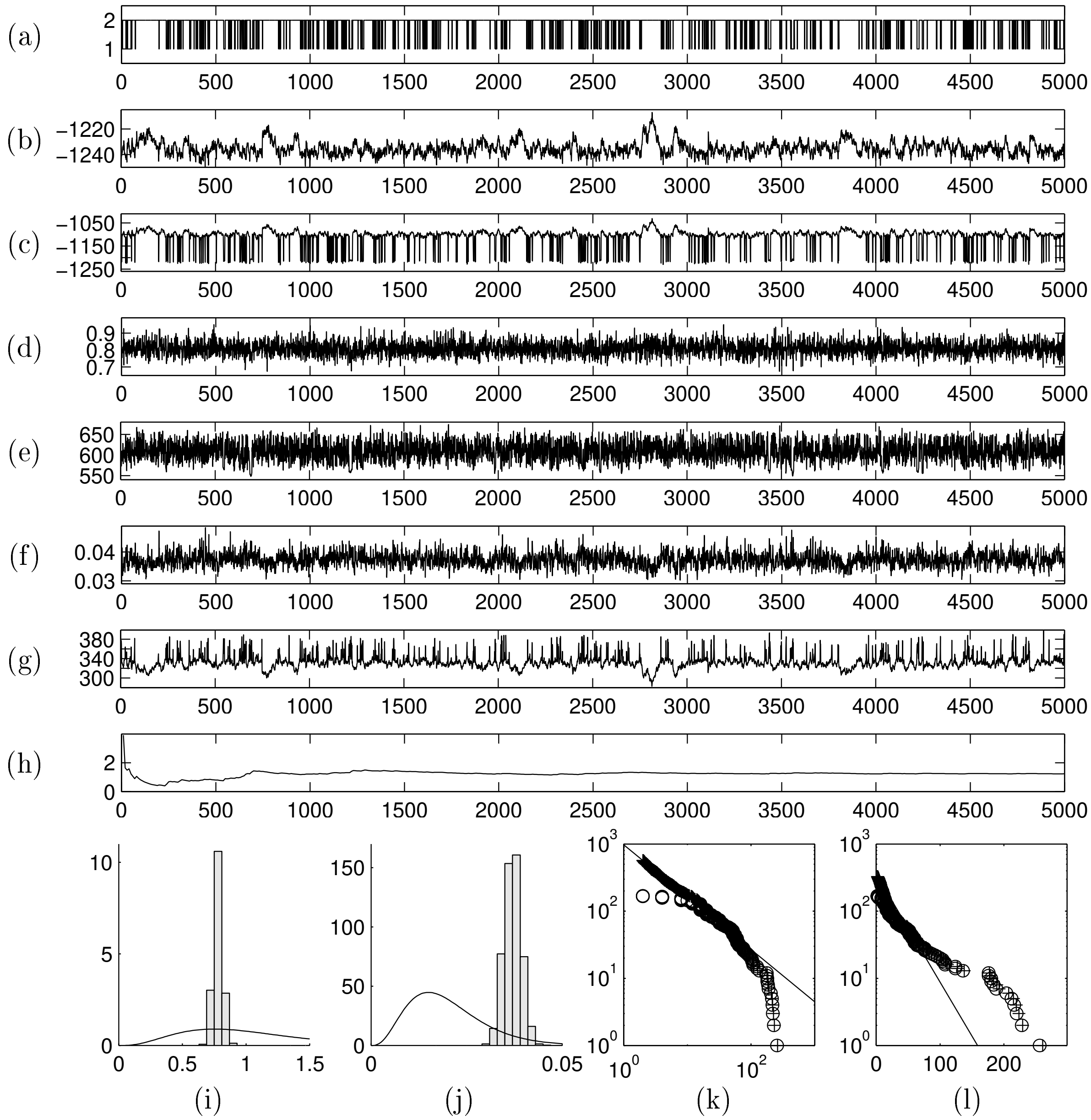


Figure 5: Output from a run of Algorithm 1 with non-exact observations. 5000 samples of (M, ψ_1, ψ_2) are generated from $\pi(m, \psi_1, \psi_2|\boldsymbol{x})$, separated by 10 iterations. The first plots are trace plots of (a) m, (b) $\log \pi(m, \psi_1, \psi_2|\boldsymbol{x})$, (c) $\log h(\boldsymbol{x}|m, \psi_m)$, (d) θ_1, (e) k_1, (f) θ_2 and (g) k_2 and (h) $\hat{B}_{12}(\boldsymbol{x})$. Histograms of (i) $\{\theta_1|m = m_1\}$ and (j) $\{\theta_2|m = m_2\}$ are plotted, comparing with the corresponding prior pdfs (solid lines). The last two figures shows $\hat{k}_m S(y|m, \hat{\theta}_m)$ (solid lines), see Expression (45), where $\hat{k}_m$ is the mean value of the dimension K_m obtained in the simulations. The function is compared to empirical estimates based on observations $\boldsymbol{x}$ and a realization $\boldsymbol{y}$ from $f(\boldsymbol{y}|\boldsymbol{x}, m, \hat{\psi}_m)$. The functions are plotted on log-log scale under model 1, figure (k), and with logarithmic vertical axis under model 2, figure (l).

is not surprising, since 5000 realizations of Ψ_m is now used to estimate a two-dimensional joint distribution, while in the case of exact observations a univariate pdf was estimated based on the same number of samples. The mismatch between $\tilde{B}_{12}(\boldsymbol{x})$ and $B_{12}(\boldsymbol{x})$ gives an acceptance probability $\alpha(m'|m)$ far from 1, and instead of a negative autocorrelation a strong positive autocorrelation is obtained. The quality of the estimate $\tilde{B}_{12}(\boldsymbol{x})$ and the pseudopriors $\pi(\psi_m|m_*)$ could possibly be improved by generating a larger number of samples of Ψ_1 and Ψ_2 in the initial run of the McMC algorithm. However, this would require longer time for each iteration in the final run of the algorithm, since the time consumption for evaluating the pseudoprior in Expression (25) increases with the number of samples ψ_m^j from the initial run.

Although no negative autocorrelation in M is obtained, Figure 5a shows that the Markov chain still alternates satisfactorily between m_1 and m_2. The value of the target pdf in Expression (42), Figure 5b, the likelihood function (40), Figure 5c, and the samples of the parameters Ψ_1 and Ψ_2, Figure 5d–g, all indicate that convergence of the Markov chain is reached. From Figure 5h it seems that 50 000 iterations is sufficient to give a reliable estimate of Bayes factor. Figure 5k,l shows realizations of $\boldsymbol{Y}$ from the posterior pdfs $f(\boldsymbol{y}|\boldsymbol{x}, m, \hat{\psi}_m)$, comparing the estimated function $S(y|m, \hat{\theta}_m)$, Expression (45), with the empirical estimates obtained based on observations $\boldsymbol{x}$ and samples $\boldsymbol{y}$. The empirical functions are multiplied by the number of faults, n and $\hat{k}_m$ for $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively, to illustrate at what scales the unobserved variables Y_j are sampled.

6 Conclusion

A model choice problem is considered, where inference is made based on observations affected by sampling errors. The underlying unknown variables can be divided into a high-dimensional part with an analytically tractable posterior distribution and a low-dimensional model parameter with an analytically non-tractable posterior. An McMC algorithm is constructed to estimate Bayes factor and make inference about the low-dimensional model parameters simultaneously. The unknown high-dimensional variable can afterwards be sampled from its posterior distribution. In order to obtain a completely specified target distribution of the McMC algorithm, the concept of pseudopriors introduced by Carlin and Chib (1995) is adopted. The freedom of choice for these pseudopriors is explored, under the major goal of reducing uncertainty in the estimated Bayes factor.

The performance of the algorithm is studied for a situation where Bayes factor is analytically tractable, and all estimates can be compared to the corresponding true, known values. The algorithm is observed to give good results in this case, where the observations are assumed to be exact and the model parameter is one dimensional. In this case a strong negative correlation between succeeding realizations of the model indicator can be produced, reducing the uncertainty in the estimated Bayes factor.

An example of non-exact data is provided. It is observed that as the dimension of the unknown parameters included in the McMC algorithm increases, it becomes more difficult to find good pseudopriors based on a relatively small number of samples of the parameters. This results in a slower mixing of M, and no negative autocorrelation is obtained. However, the model indicator M still changes relatively often throughout the simulations, and the algorithm seems to perform satisfactory also in this case.

If the posterior distribution of the unobserved high-dimensional parameter, or parts of it, were not analytically tractable, this variable could also have been included in the sampling algorithm. However, this expansion of the sample space is assumed to reduce the efficiency of the algorithm considerably.

Acknowledgements

Hilde G. Borgos is funded by a PhD grant from VISTA, a research corporation between The Norwegian Academy of Science and Letters and Den norske stats oljeselskap a.s. (Statoil).

References

Ackermann, R. V., Withjack, M. O. and Schlische, R. W. (1999). The geometric and statistical evolution of normal fault systems: An experimental study of the effects of mechanical layer thickness on scaling laws, submitted to *Journal of Structural Geology*.

Aitkin, M. (1991). Posterior Bayes Factors, *J. R. Statist. Soc. B* **53**(1): 111–142.

Berger, J. O. and Pericchi, L. R. (1996). The Intrinsic Bayes Factor for Model Selection and Prediction, *Journal of the American Statistical Association* **91**(433): 109–122.

Besag, J., Green, P., Higdon, D. and Mengersen, K. (1995). Bayesian Computation and Stochastic Systems (with discussion), *Statistical Science* **10**(1): 3–66.

Borgos, H. G., Omre, H. and Townsend, C. (2000). Model choice for fault size distribution, *Preprint Statistics 3/2000*, Department of Mathematical Sciences, Norwegian University of Science and Technology, Trondheim, Norway.

Carlin, B. P. and Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods, *J. R. Statist. Soc. B* **57**(3): 473–484.

Carlin, P. and Polson, G. (1991). Inference for nonconjugate Bayesian Models using the Gibbs sampler, *The Canadian Journal of Statistics* **19**(4): 399–405.

Chib, S. (1995). Marginal Likelihood From the Gibbs Output, *Journal of the American Statistical Association* **90**(432): 1313–1321.

Cowie, P. A., Malinverno, A., Ryan, W. B. F. and Edwards, M. H. (1994). Quantitative fault studies on the East Pacific Rise: A comparison of sonar image techniques, *Journal of Geophysical Research* **99**(B8): 15205–15218.

Dellaportas, P., Forster, J. J. and Ntzoufras, I. (1998). On Bayesian Model and Variable Selection Using MCMC, *Technical Report No 40*, Department of Statistics, Athens University of Economics and Business. Submitted to *Statistics and Computing*.

DiCiccio, T. J., Kass, R. E., Raftery, A. and Wasserman, L. (1997). Computing Bayes Factors By Combining Simulation and Asymptotic Approximations, *Journal of the American Statistical Association* **92**(439): 903–915.

Evans, M. and Swartz, T. (1995). Methods for Approximating Integrals in Statistics with Special Emphasis on Bayesian Integration Problems, *Statistical Science* **10**(3): 254–272.

Fossen, H. and Rørnes, A. (1996). Properties of fault populations in the Gullfaks Field, northern North Sea, *Journal of Structural Geology* **18**(2/3): 179–190.

Gelman, A. and Meng, X.-L. (1998). Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling, *Statistical Science* **13**(2): 163–185.

George, E. I. and McCulloch, R. E. (1993). Variable Selection Via Gibbs Sampling, *Journal of the American Statistical Association* **88**(423): 881–889.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination, *Biometrica* **82**(4): 711–732.

Han, C. and Carlin, B. P. (2000). MCMC Methods for Computing Bayes Factors: A Comparative Review, *Research Report 2000-001*, Division of Biostatistics, University of Minnesota.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications, *Biometrica* **57**: 97–109.

JSG (1996). *Journal of Structural Geology*, Vol. 18, No. 2/3, Special Issue, Scaling Laws for Fault and Fracture Populations – Analyses and Applications.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors, *Journal of the American Statistical Association* **90**(430): 773–795.

O'Hagan, A. (1991). Contribution to the discussion of "Posterior Bayes Factors" (by M. Aitkin), *J. R. Statist. Soc. B* **53**(1): 136.

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison, *J. R. Statist. Soc. B* **57**(1): 99–138.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Spryopoulos, C., Griffith, W. J., Scholz, C. H. and Shaw, B. E. (1999). Experimental evidence for different strain regimes of crack populations in a clay model, *Geophysical Research Letters* **26**(8): 1081–1084.